Machine learning: lecture 23

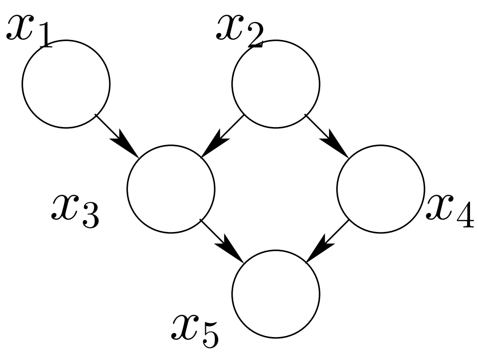
Tommi S. Jaakkola
MIT CSAIL
tommi@csail.mit.edu

Outline

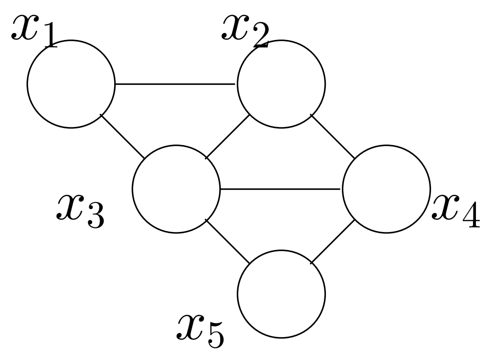
- Exact inference (quickly)
  – message passing in junction trees
- Approximate inference
  – belief propagation
  – sampling
- Review for the final
  – what is important, what is not

Exact inference: key steps

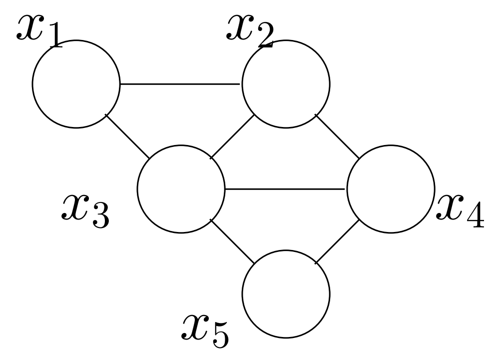
- Baysian network, moralization, triangulation



original graph

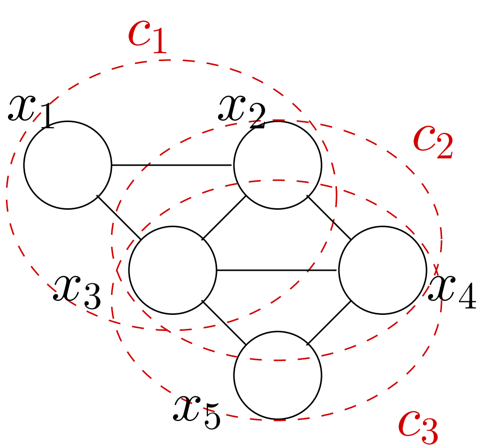


moral graph

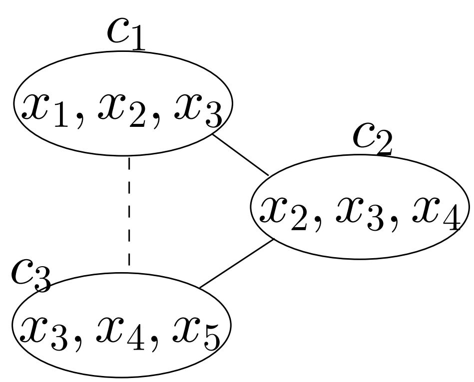


already triangulated

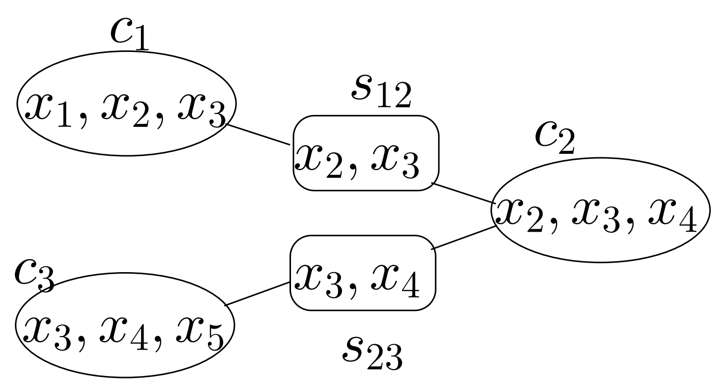
- Cliques, clique graph, and junction tree



cliques



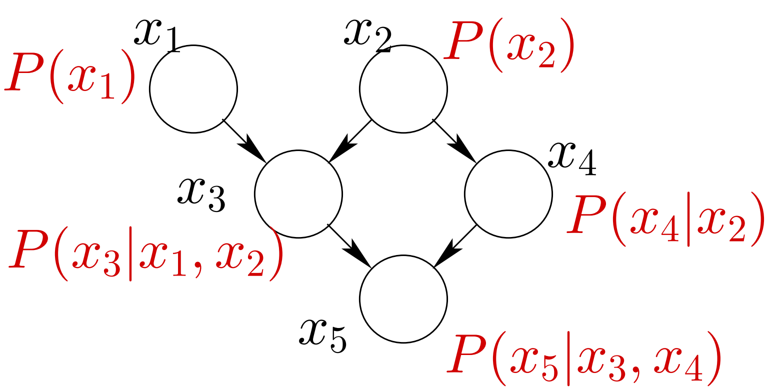
clique tree

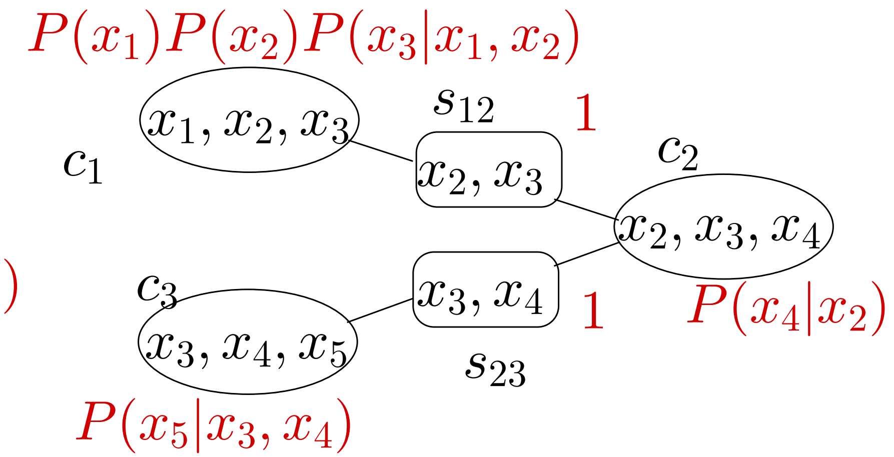


junction tree

Exact inference: potentials

- Associating graphs and potentials



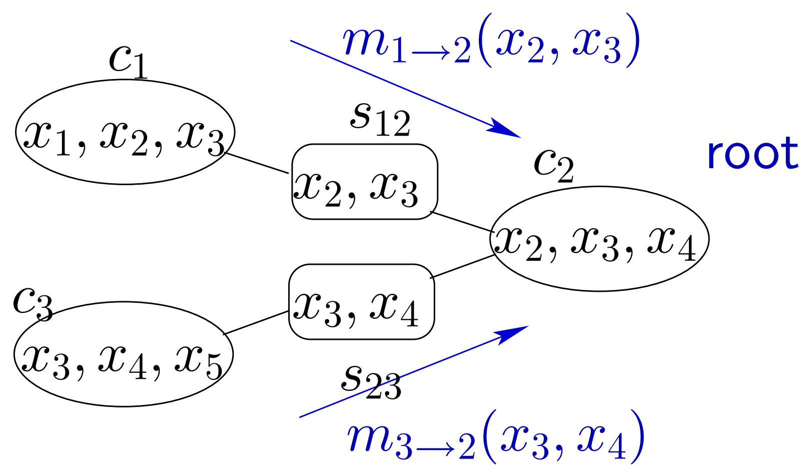


original graph w/ probabilities　　　junction tree w/ probs

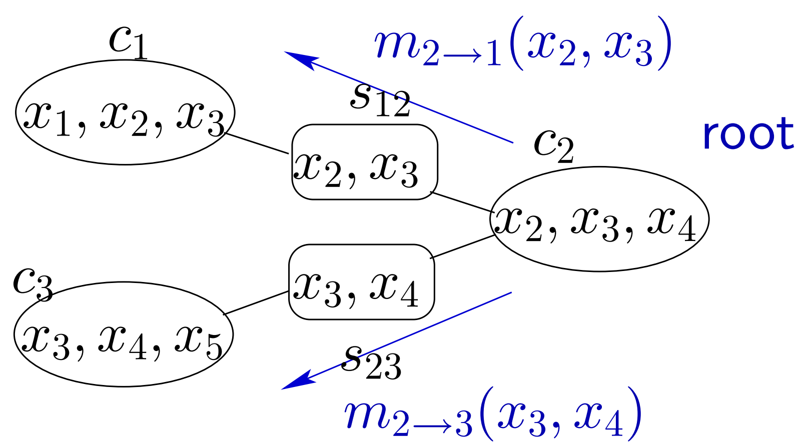
$$P(x_1,\ldots,x_5) = \frac{\psi_{c_1}(x_1,x_2,x_3)\psi_{c_2}(x_2,x_3,x_4)\psi_{c_3}(x_3,x_4,x_5)}{\psi_{s_{12}}(x_2,x_3)\psi_{s_{23}}(x_3,x_4)}$$

Exact inference: message passing

- Select a root clique
- Collect evidence

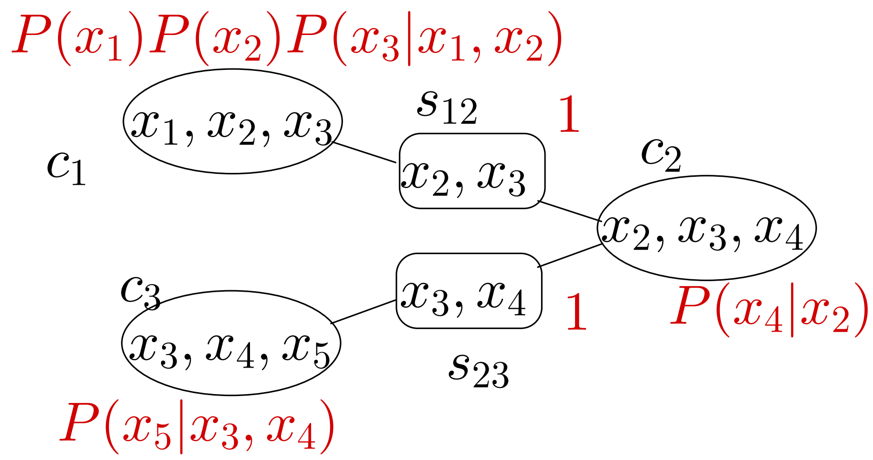


- Distribute evidence



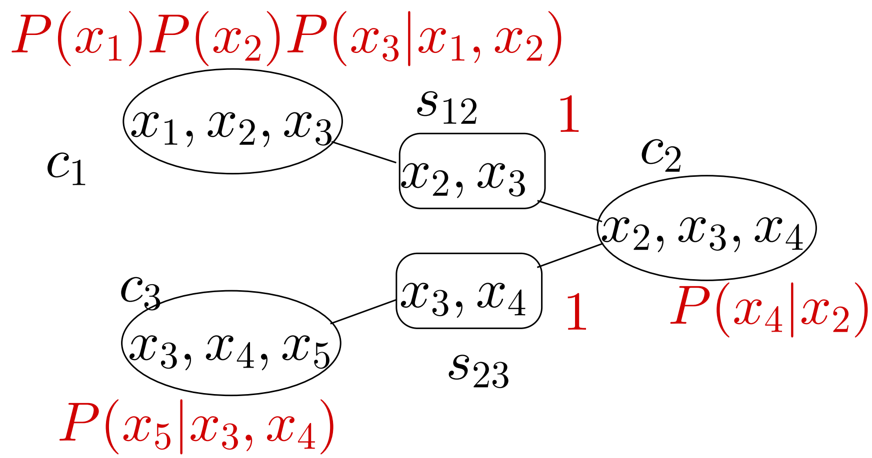
Exact inference: message passing

- Collect evidence



Exact inference: message passing

- Collect evidence



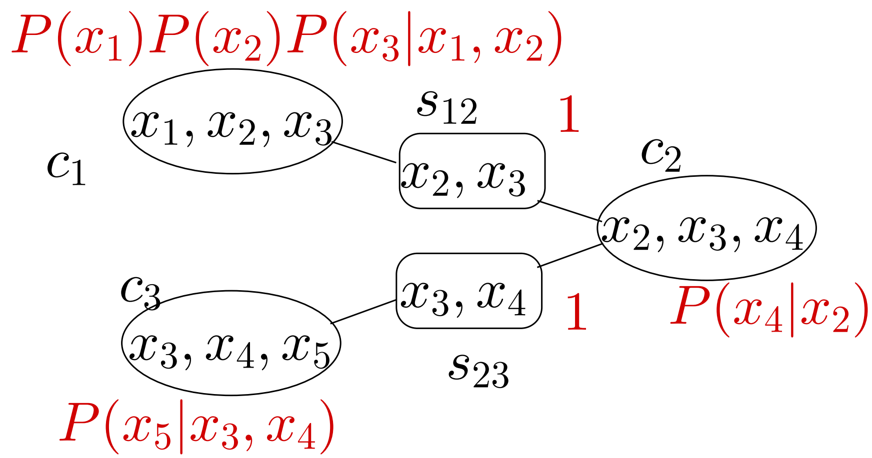
Evaluate new separators:

$$\psi'_{s_{12}}(x_2, x_3) = \sum_{x_1} \psi_{c_1}(x_1, x_2, x_3) = P(x_2, x_3)$$

$$\psi'_{s_{23}}(x_3, x_4) = \sum_{x_5} \psi_{c_3}(x_3, x_4, x_5) = 1$$

Exact inference: message passing

- Collect evidence



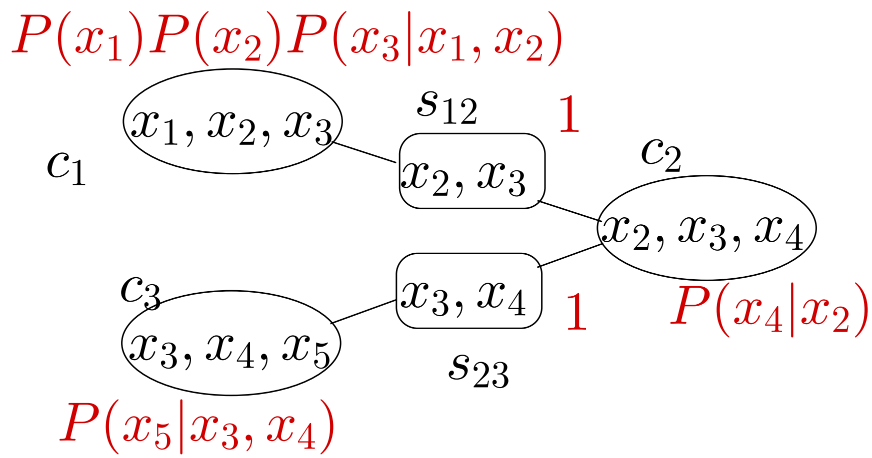
Messages (not explicitly used in the algorithm):

$$m_{1\to 2}(x_2, x_3) = \frac{\psi'_{s_{12}}(x_2, x_3)}{\psi_{s_{12}}(x_2, x_3)} = \frac{P(x_2, x_3)}{1}$$

$$m_{3\to 2}(x_3, x_4) = \frac{\psi'_{s_{23}}(x_3, x_4)}{\psi_{s_{23}}(x_3, x_4)} = \frac{1}{1}$$

Exact inference: message passing

- Collect evidence



Update clique potentials (based on messages):

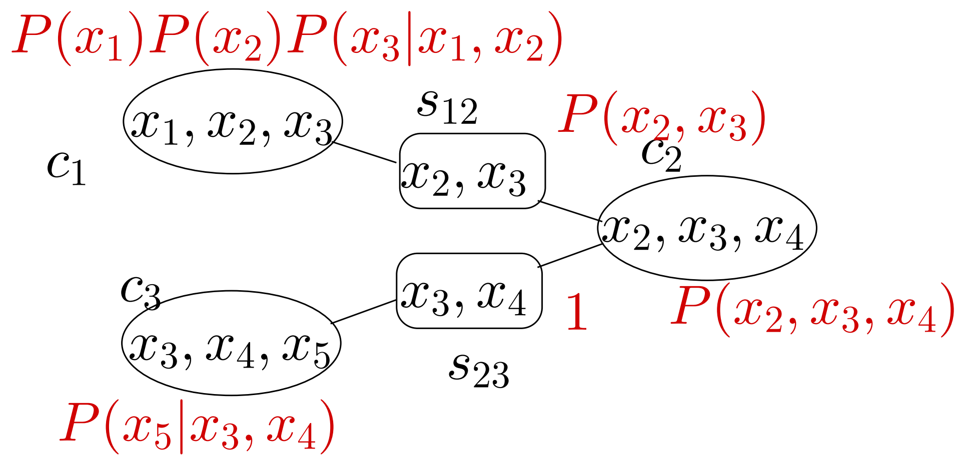
$$\psi_{c_2}(x_2,x_3,x_4) \;\leftarrow\; \underbrace{\frac{\psi'_{s_{12}}(x_2,x_3)}{\psi_{s_{12}}(x_2,x_3)}}_{m_{1\to 2}(x_2,x_3)} \cdot \underbrace{\frac{\psi'_{s_{23}}(x_3,x_4)}{\psi_{s_{23}}(x_3,x_4)}}_{m_{3\to 2}(x_3,x_4)} \cdot \psi_{c_2}(x_2,x_3,x_4)$$

$$= P(x_2,x_3)\cdot 1 \cdot P(x_4|x_2) = P(x_2,x_3,x_4)$$

followed by $\psi_{s_{12}} \leftarrow \psi'_{s_{12}}$ and $\psi_{s_{23}} \leftarrow \psi'_{s_{23}}$

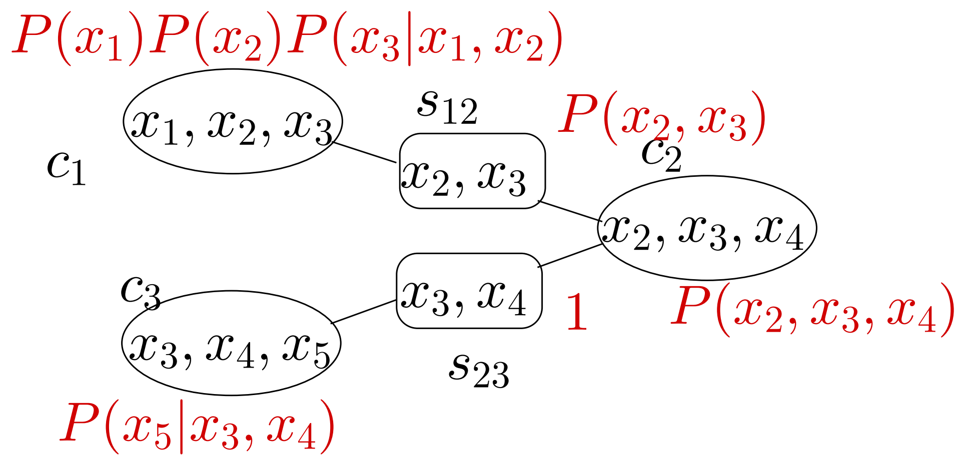
Exact inference: message passing

- Distribute evidence



Exact inference: message passing

- Distribute evidence



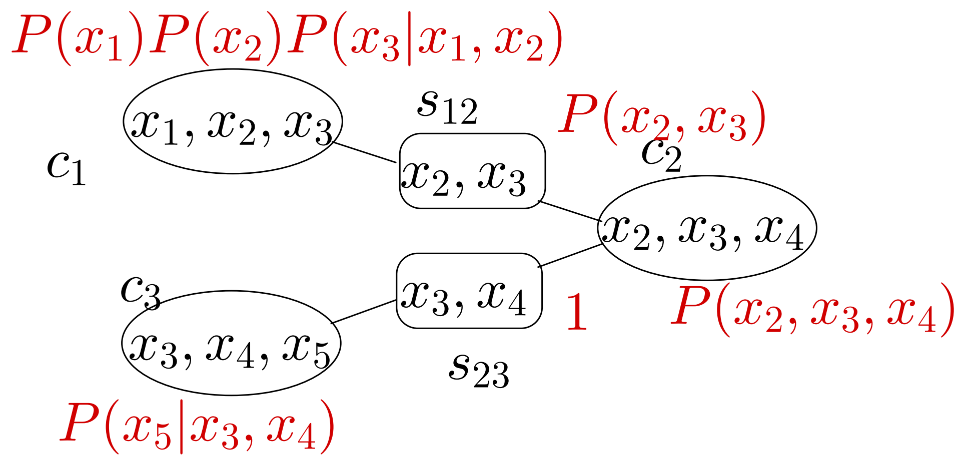
Evaluate new separators:

$$\psi'_{s_{12}}(x_2, x_3) = \sum_{x_4} \psi_{c_2}(x_2, x_3, x_4) = P(x_2, x_3)$$

$$\psi'_{s_{23}}(x_3, x_4) = \sum_{x_2} \psi_{c_2}(x_2, x_3, x_4) = P(x_3, x_4)$$

Exact inference: message passing

- Distribute evidence



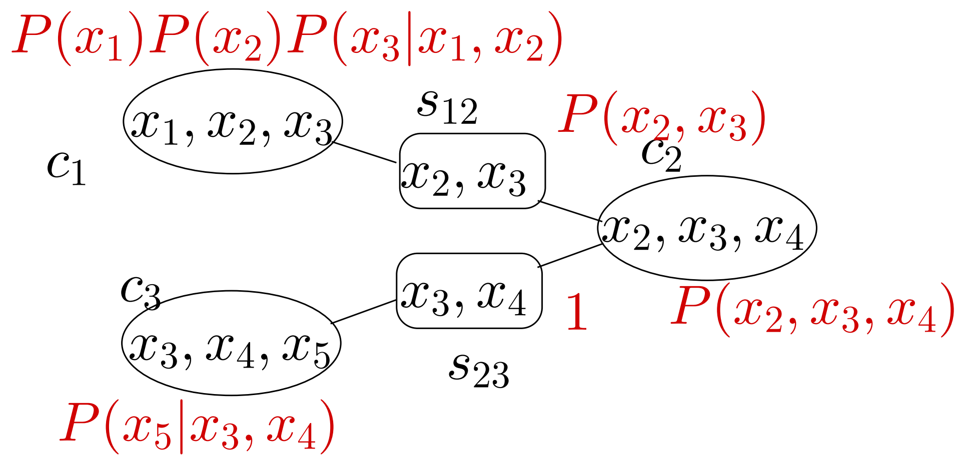
Messages (not explicitly used in the algorithm):

$$m_{2\to1}(x_2, x_3) = \frac{\psi'_{s_{12}}(x_2, x_3)}{\psi_{s_{12}}(x_2, x_3)} = \frac{P(x_2, x_3)}{P(x_2, x_3)} = 1$$

$$m_{2\to3}(x_3, x_4) = \frac{\psi'_{s_{23}}(x_3, x_4)}{\psi_{s_{23}}(x_3, x_4)} = \frac{P(x_3, x_4)}{1}$$

Exact inference: message passing

- Distribute evidence



Update clique potentials (based on messages):

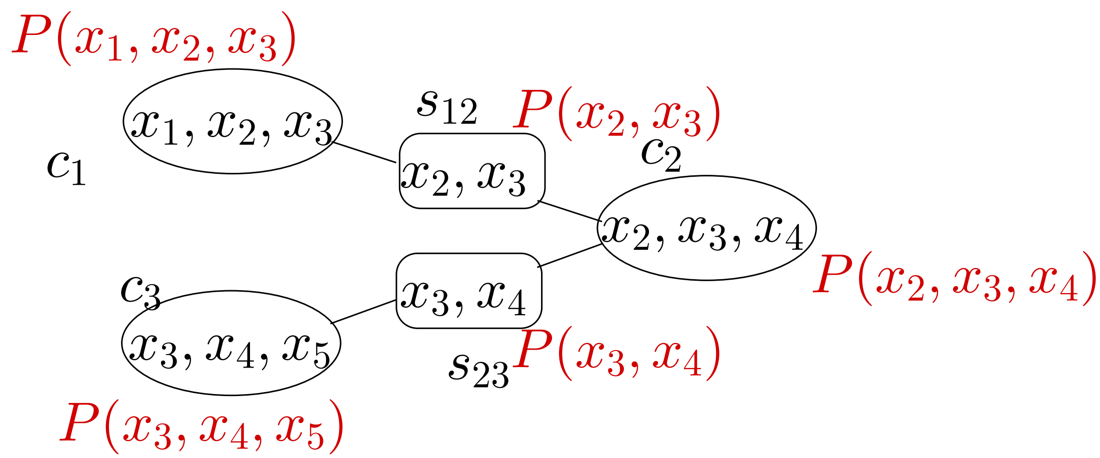
$$\psi_{c_1}(x_1,x_2,x_3) \;\leftarrow\; \frac{\psi'_{s_{12}}(x_2,x_3)}{\psi_{s_{12}}(x_2,x_3)}\psi_{c_1}(x_1,x_2,x_3) \;=\; P(x_1,x_2,x_3)$$

$$\psi_{c_3}(x_3,x_4,x_5) \;\leftarrow\; \frac{\psi'_{s_{23}}(x_3,x_4)}{\psi_{s_{23}}(x_3,x_4)}\cdot\psi_{c_3}(x_3,x_4,x_5) \;=\; P(x_3,x_4,x_5)$$

followed by $\psi_{s_{12}} \leftarrow \psi'_{s_{12}}$ and $\psi_{s_{23}} \leftarrow \psi'_{s_{23}}$

Exact inference

- After the collect and distribute steps the marginal probabilities are stored *locally* at the clique potentials (and the separators)



$$P(x_1, \ldots, x_5) = \frac{P(x_1, x_2, x_3)P(x_2, x_3, x_4)P(x_3, x_4, x_5)}{P(x_2, x_3)P(x_3, x_4)}$$

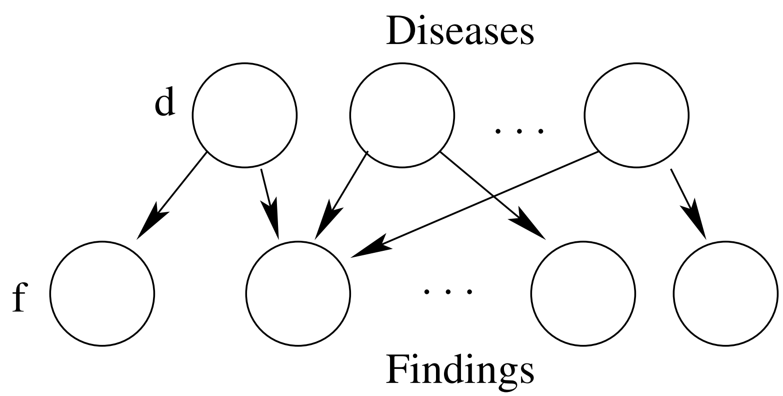
More generally, the resulting potentials would be proportional to the posterior marginals, e.g., $P(x_1, x_2, x_3, \text{data})$, which can be easily normalized.

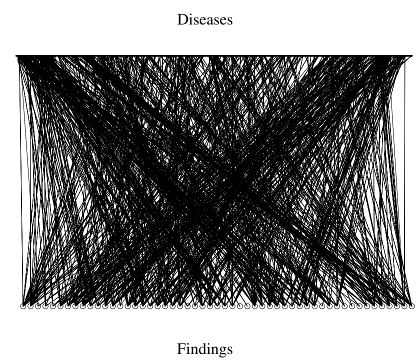
Outline

- Exact inference (quickly)
  - message passing in junction trees
- Approximate inference
  - belief propagation
  - sampling
- Review for the final
  - what is important, what is not

Approximate inference: motivation

- We cannot solve the medical diagnosis problem(s) with exact inference algorithms

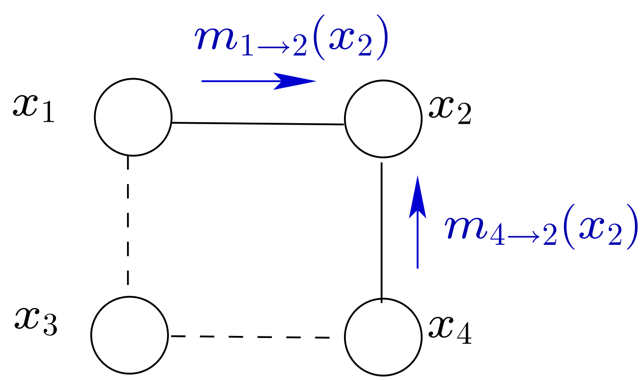




  – the largest clique has over 100 variables (the corresponding potential function or table would involve more than 2^{100} elements)

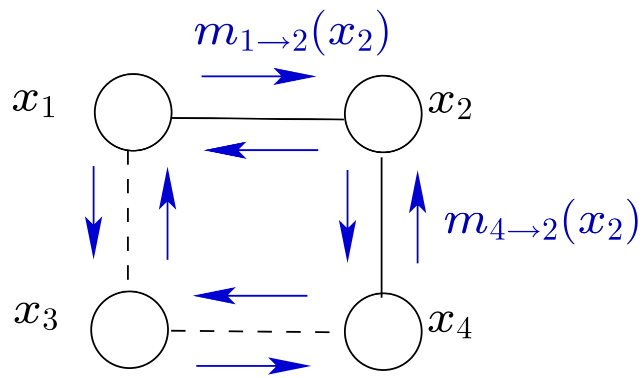
Approximate inference: belief propagation

- The message passing algorithm is appropriate when the model is a (clique) tree
  - we need a unique path of influence between any (sets of) variables
- We can still apply the message passing algorithm even if the model is not a tree (message passing operations are defined locally)



  - convergence?
  - accuracy?

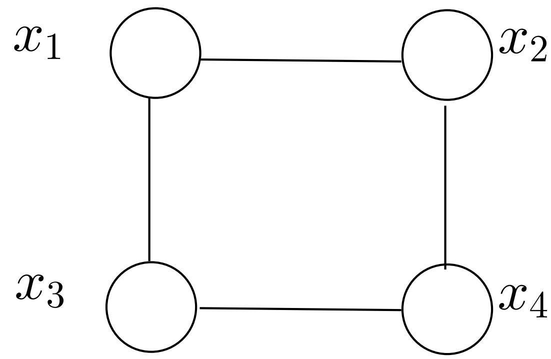
Approximate inference: belief propagation



- a set of locally consistent messages (fixed point of the algorithm) always exists
- the accuracy of the resulting marginals related to the length of the shortest cycle
- stronger guarantees exist for finding most likely configurations of variables

- Works well in many large scale applications
  - decoding turbo (and other) codes, image processing, molecular networks, protein structure, etc.

Approximate inference: sampling



- If we could draw samples $\mathbf{x}^t = \{x_1^t, x_2^t, x_3^t, x_4^t\}$ from $P(\mathbf{x})$, we could easily and accurately evaluate any marginals

$$P(x_1 = 0) \approx \frac{1}{T} \sum_{t=1}^{T} \delta(x_1^t, 0)$$

where $\delta(x_1^t, 0) = 1$ whenever $x_1^t = 0$ and zero otherwise.

- But it is hard to draw samples...

Simple remedy: importance sampling

- We can instead draw samples from a much simpler distribution $Q(\mathbf{x})$ (e.g., where the variables may be independent) and evaluate marginals according to

$$
\begin{aligned}
P(x_1 = 0) &= \sum_{\mathbf{x}} P(\mathbf{x})\delta(x_1, 0) \\
&= \sum_{\mathbf{x}} Q(\mathbf{x})\frac{P(\mathbf{x})}{Q(\mathbf{x})}\delta(x_1, 0) \\
&\approx \frac{1}{T}\sum_{t=1}^{T} \frac{P(\mathbf{x}^t)}{Q(\mathbf{x}^t)}\delta(x_1^t, 0)
\end{aligned}
$$

where the samples $\mathbf{x}^t$ are now drawn from $Q(\mathbf{x})$.

- But the resulting marginals may not even lie in $[0, 1]$...

Likelihood weighted sampling

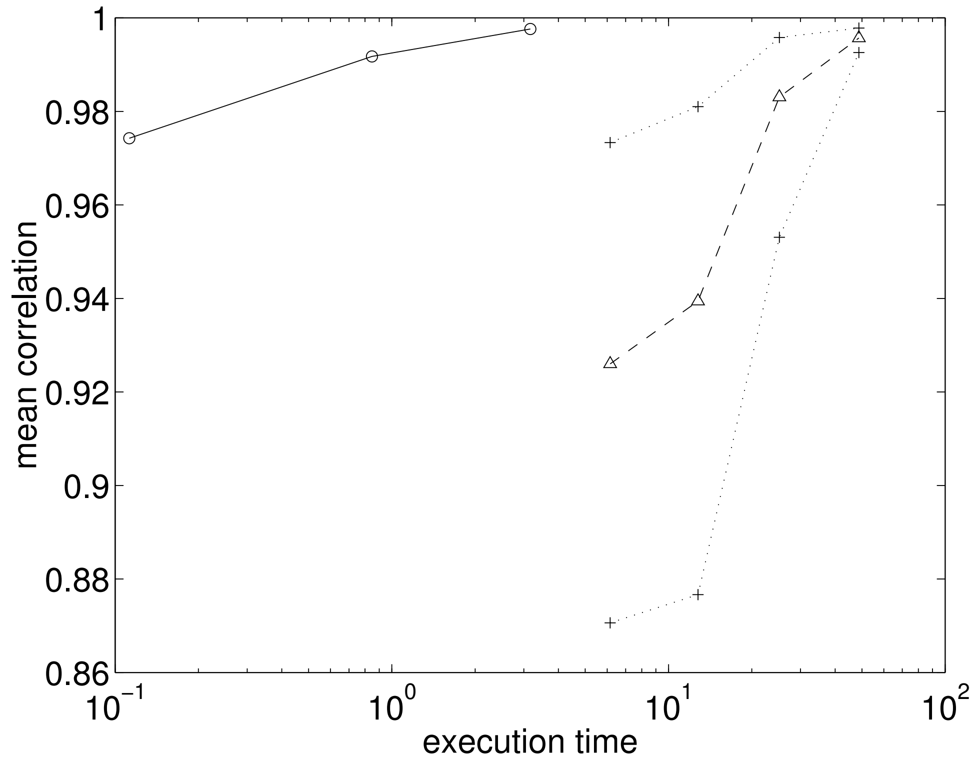
- A better (but biased) sampling approximation is given by a likelihood weighted average

$$P(x_1 = 0) \approx \frac{\frac{1}{T}\sum_{t=1}^{T} \frac{P(\mathbf{x}^t)}{Q(\mathbf{x}^t)}\delta(x_1^t, 0)}{\frac{1}{T}\sum_{t=1}^{T} \frac{P(\mathbf{x}^t)}{Q(\mathbf{x}^t)}}$$

- Any factored sampling distribution $Q(\mathbf{x}) = \prod_i Q_i(x_i)$ can be adjusted adaptively on the basis of the marginals computed so far

Back to medical diagnosis problem

- Likelihood weighted sampling works... sort of



The figure shows the overall correlation between the estimated and exact posterior marginals (in simple cases)

Outline

- Exact inference (quickly)
  - message passing in junction trees
- Approximate inference
  - belief propagation
  - sampling
- Review for the final
  - what is important, what is not

The final

- General points
  - exam is comprehensive, not limited to the second half
  - emphasis on concepts, integration

The final

- Major topics
  - regression and classification, additive models
  - discriminative and generative classifiers
  - estimation, over-fitting, generalization
  - regularization, support vector machines
  - feature selection, boosting
  - complexity, compression, model selection
  - mixtures, EM, conditional mixtures
  - clustering formulations, methods
  - HMMs, algorithms, modeling
  - Bayesian networks, graph, inference